Response to Comments by Anonymous Referee #2

We thank the reviewer for taking the time to review the manuscript and provide comments. We agree that there are areas that need clarifications. Suggestion for changes in figures was also much appreciated. The reviewer comments are highlighted in bold with our responses written below. We also state the changes that will be made in the revised manuscript with regard to each comment.

1) **The last sentence of the abstract states "This approach offers a promising prospect of using physics-based machine learning applications to other instruments." However, as this algorithm did not originate in this study or for the OMI instrument, and so has already been demonstrated for a number of other instruments, as phrased here the statement doesn't seem valid. I would suggest the sentence is either rephrased (if that wasn't it's intended meaning), or removed.**

We agree that the statement is not necessary and have removed it from the abstract text.

2) **The section describes IASI retrievals of SO2 height, and follows it with the phrase ('For these techniques, extensive radiative transfer modeling is needed: : :). However, in the Clarisse 2014 paper referenced, this describes a fast retrieval scheme, where this statement may not follow. I would suggest checking and amending the text appropriately.**

It is true that the 2014 paper discusses an updated IASI retrieval that is much faster, therefore the 2014 reference should not be included after that sentence. We have included another sentence that highlights the retrievals from the 2014 paper separately.

3) **Section 2.4 Line 264-266: "The output is a predicted SO2 layer height based on the input of a radiance spectra and associated parameters, including VZA, SZA, RAA, surface albedo and surface pressure, for a single OMI pixel." Is this sentence in the right place – I found the flow of the paragraph a bit confusing, as it then jumps back to talking about convolving the irradiance spectra and then applying PCA?**

Yes, we agree that sentence was out of place and a bit redundant. It has been removed from text.

4) **Section 2.4, Line 278: The text assumes that readers will be familiar with the OMI row anomaly, which may not be the case – it would be useful to explain this somewhere.**

We acknowledge that we did not provide background information on the row anomaly. We added an introduction to the row anomaly in Section 1 when introducing the OMI instrument. The cause of the anomaly is also indicated.

5) **Section 3: Tables 2 and 3 need more explanation in the text here e.g. RMSE is mentioned, but it's not explicitly stated what this represents anywhere. Also, from**

their captions, I would have expected the RMSE numbers in Table 2 for VCD > 40 and SZA
< 75, to be the same as the RMSE in Table 3 for SNR = 1000. However, the numbers don't agree. What is the reason for the difference?

- Thank you for pointing this out. It seems that the noise analysis was performed earlier with different neural network conditions, as those were changed multiple times when trying to optimize the training. We have redone analysis for Table 3 (below) using the same NN setup and test dataset as for the Table 2 which has resolved the discrepancy.

Table 3: The RMSE and the mean absolute difference of all data points in the independent test set after adding noise as indicated by different SNR values. All other parameters and input data were kept constant. SZA < 75 degrees and SO2 VCD > 40 DU were excluded from the test set for these comparisons.

	No noise	SNR=1000	750	500	200	100
Mean Absolute Difference (y_known - y_pred) (km)	0.894	0.904	0.939	0.996	1.114	1.362
RMSE (km)	1.454	1.498	1.521	1.632	1.807	2.143
R-coefficient	0.988	0.985	0.983	0.980	0.972	0.955

- On the second point, we will be sure to explain Table 2 and 3 better in the text. The RMSE was a metric used to evaluate neural network performance, more specifically the error difference between the "predicted" height (i.e. output from NN based on test data) and "actual" height which is the output from the training set that was used in training.

6) Section 4: What are the expected uncertainties of the validation data products used - the text talks about reasonable agreement, but there are differences of several km's in some cases, so it would be useful to know if that can be explained by uncertainties in the other datasets as well? In particular, for Kasatochi, the quoted values for prior OMI retrievals are a few km's lower – has the reason for this been looked at in more detail?

For TROPOMI retrieval (Hedelt et al., 2019) there is a stated retrieval uncertainty of < 2km for SO_2 column of greater than 20 DU. However, this is only for the retrieval using the synthetic data. Using real data also adds a certain degree of error. The IASI retrieval also contains an uncertainty of 2 km. In some cases there is more than 2 km difference between different datasets. In addition to uncertainties within validation datasets there are also differences between retrieval technique which could also add to the differences.

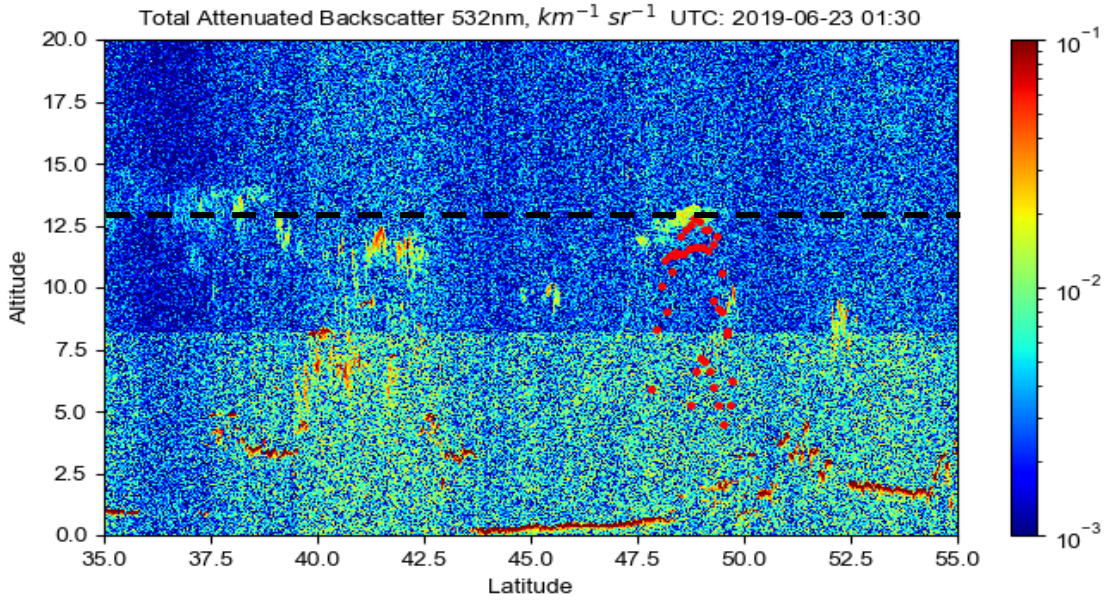
The previous OMI SO_2 height retrieval pertaining to Kasatochi is from Yang et al, 2010. The height values for Kasatochi were found to be around 9-11 km with uncertainty of up to 2 km as well. We believe this is within reasonable range of our retrieval, especially taking the difference

of retrieval technique into account. We agree that this is important to discuss, and we have added a few sentences in the discussion clarifying this.

7) **Figures 4, 5, 7: These would be clearer if all the instruments were plotted on the same colour scale and lat/lon range. Also, if possible, replot the Caliop data to focus on the relevant region. Similarly for Figure 8, it would be clearer if all the instruments were plotted on the same axes.**

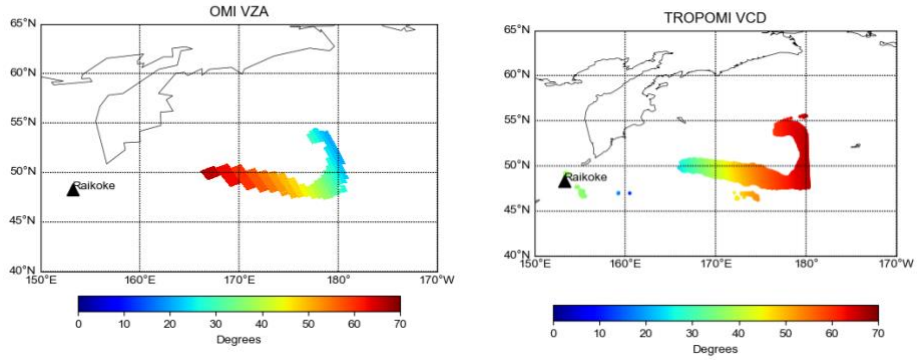
Thank you for pointing this out. We have replotted figures so color scales and coordinate ranges match. Figure 4c has also been updated using GOME-2 data.

For the CALIPSO lidar data, we agree that it would be helpful to focus in on the region with the volcanic plume. The figures have been replotted for the revised manuscript. As an example here is one for Raikoke (June 23rd, 2019) which would correspond to Figure 9a.



8) **In Figures 7 and 8 for the Raikoke eruption, the distribution of values for OMI and TROPOMI seem to be mirrored e.g. OMI has a tail of lower values while TROPOMI has a tail of higher values. Is there an explanation for this?**

That is a good observation. It would be difficult to pin point the exact reason because of instrument and retrieval difference, but one interesting observation was that radiance measurement from the instruments were obtained from different cross track positions (rows).In other words the retrievals for OMI were on the left side of the swath (high to low VZA) while for TROPOMI it was the opposite side of the its swath (low to high). Since VZA was one of the parameters involved in training process, this could have some effects on the mirrored distribution. The maps below show VZA plotted for the SO_2 plume area.



However, we do not think that there should be a big dependence in the retrieval on VZA. Another explanation could be differences in signal to noise ration (SNR) at nadir (VZA ~ 0) versus the edges where there can be some degradation. There are also differences in SNR between OMI and TROPOMI and furthermore, TROPOMI spectra in the UV is affected by instrument degradation issues. We have included a few sentences about this in Section 4 that discuss these points.

9) Section 4.3 – the first paragraph reads as an introduction to section 4.2 too – should the ordering be changed?

That is a good catch. This paragraph was moved to the beginning of section 4.2 instead.

10) Section 4.5 'Discussion of errors'. Have the authors looked in any more detail on the impact of some of their assumptions in the radiative transfer modelling on the retrieval errors? E.g. they mention that using a fixed solar irradiance spectrum will be less accurate than using the OMI solar measurements. Has the expected impact on this been quantified? Is a fuller assessment of these sorts of errors planned as part of their future work?

Yes, quantifying the effect of instrument versus the fixed spectrum is more for future work. In general it makes sense that using the irradiance measurements from the same instrument would carry less potential error, however there are benefits for using a fixed reference spectrum. The downside in using instrument irradiance is that the solar irradiance spectrum varies month to month and for each row. This would require far more computation and additional complexity in training thus making it less suitable for future NRT implementations. Additionally we consider applying this algorithm to other instruments, which with a fixed irradiance spectrum is possible without having to repeat the calculations of synthetic radiances. We will remove the accuracy comparison from the manuscript since it is not backed up quantitatively at the moment

There are also many aspects of both the radiative transfer modeling and neural network that can be explored more in depth (i.e. sensitivity analyses). Our primary goal was to obtain a robust algorithm and reasonably accurate algorithm first, however we certainly plan to explore certain sensitivities as future work and if they make a significant impact on result.

11) Why are the figures 1A and 2A supplemental, as they're directly referenced in the text?

We initially placed them in supplemental section since they are finer details of the methodology. However, since they ended up being referenced, we will move them to main figures and adjust figure numberings accordingly.

12) Conclusion: Line 472 "with absolute errors of up to 1.5km" – This seems to be the first time that number is quoted, and given the uncertainties and difficulties comparing instruments it may be too strong to put a hard number on an absolute error e.g. The section on errors just mentions 1-2km differences. I think if it's quoted like this, it would be good to back it up with more quantitative information as to where it came from. Otherwise, I would rephrase. (Similarly, the abstract quotes errors of 1-1.5 km's, which should also be made consistent).

We changed the statement to "1-2 km" since this is the likely range of errors. You are correct in pointing out that stating an exact error value is not valid here.